Semantic Answer Validation in Question Answering Systems for Reading Comprehension Tests

Helena Gómez-Adorno, David Pinto, Darnes Vilariño

Faculty of Computer Science
Benemérita Universidad Autónoma de Puebla
Av. San Claudio y 14 Sur,C.P. 72570, Puebla, Mexico
{helena.gomez,dpinto,darnes}@cs.buap.mx,
http://www.cs.buap.mx/

Abstract. In this paper it is presented a methodology for tackling the problem of answer validation in question answering for reading comprehension tests. The implemented system accepts a document as input and it answers multiple choice questions about it based on semantic similarity measures. It uses the Lucene information retrieval engine for carrying out information extraction employing additional automated linguistic processing such as stemming, anaphora resolution and part-of-speech tagging. The proposed approach validates the answers, by comparing the text retrieved by Lucene for each question with respect to its candidate answers. For this purpose, a validation based on semantic similarity is executed. We have evaluated the experiments carried out in order to verify the quality of the methodology proposed using a corpus widely used in international forums. The obtained results show that the proposed system selects the correct answer to a given question with a percentage of 12% more than with a lexical similarity based validation.

Keywords: Question answering system, reading comprehension, information retrieval, semantic similarity

1 Introduction

Reading comprehension comprises the ability of human reader for understanding the main ideas written in a text. In order to evaluate the quality of reading comprehension, there exist tests that require readers for reading a story or article and answer a list of questions about it. From the point of view of automatic evaluation of reading comprehension tests, it is needed to take advantage of the techniques developed in the framework of question answering.

In this paper we present some experiments for exploring answer validation in question answering architectures that can be applied to reading comprehension tests as an evaluation method for language understanding systems (machine reading systems). Such tests take the form of standardized multiple-choice diagnostic reading skill tests.

The main idea behind QA systems for reading comprehension tests is to answer questions based on a single document. This approach is different from that of traditional QA systems, in which they have a very large corpus for searching the requested information, which implies in some cases a very different system architecture.

The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 describes the System Architecture. Section 4 presents the evaluation results in a collection of documents of the QA4MRE task at CLEF 2011. Finally, Section 5 presents the conclusions obtained, so that it outlines some future work directions.

2 Related Work

The QA for reading comprehension tests field has been inactive for a long time, due to the lack of agreement in the way the systems evaluation should be done [1] . In 2011, and later in the 2012, the CLEF conference[1] proposed a QA task for *Machine Reading (MR) systems* evaluation called QA4MRE. The task consists of reading a document and identifying answers for a set of questions about the information that is expressed or implied in the text. The questions are written in the form of multiple choices; each question has 5 different options, and only one option is the correct answer. The detection of the correct answer is specifically designed to require various types of inference, and the consideration of prior knowledge acquired from a collection of reference documents [2, 3].

The QA4MRE task encourage the interest in this research line, because it provides a single evaluation platform for the experimentation with new techniques and methodologies towards giving a solution to this problem. In this sense we can take the systems presented in this conference as state-of-the-art work for this research field.

However there exist other research works [4–6] that also have deal with the problem of QA for reading comprehension tests in the past, unfortunately with low level of accuracy.

3 System Architecture

The proposed architecture is made up of three main modules: Document processing, Information Extraction and Answer validation. Each of these modules is described in the following subsections.

3.1 Document Processing

First we analyze the queries associated to each document, applying a Part-Of-Speech (POS) tagger in order to identify the "question keywords" (what, where,

[1] The Cross-Lingual Evaluation Forum: http://www.clef-initiative.eu

when, who, etc.), and the result is passed to the *hypothesis generation* module (this module will be explained more into detail in Section 3.2).

Afterwards, we perform anaphora resolution for the documents associated with the questions using the JavaRAP[2] system. It has been observed that applying anaphora resolution in QA systems improves the results obtained, in terms of precision [7]. Given that JavaRAP does not resolve anaphors of first-person pronouns, we added the following process for the resolution of these cases:

1. Identify the author of the document, which is usually the first name in the document. For this purpose, the Stanford POS tagger[3] was used.
2. Each personal pronoun in the first person set PRP={"I", "me", "my", "myself"} generally refers to the author.
3. Replace each term of the document that is in the PRP set, by the document author name identified in step 1.

3.2 Information Extraction

Secondly, we extract the meaningful information by means of two submodules: Hypothesis Generation and Information Retrieval.

The first submodule (Hypothesis Generation) receives as input the set of questions with their multiple choice answers, which were previously processed in the previous module. We construct what we means *h*ypothesis as the concatenation of the question with each of the possible answers. This hypothesis is intended to become the input to the Information Retrieval (IR) module, i.e., the query. In order to generate the hypothesis, first the "question keyword" is identified and subsequently replaced by each of the five possible answers, thereby obtaining five hypotheses for each question. For example, given the question: **Where** was Elizabeth Pisani's friend incarcerated?. And a posible answer: **in the Philippines**. The obtained hypothesis is: **in the Philippines** was Elizabeth Pisani's friend incarcerated.

The benefit of using these hypotheses as queries for the IR module is to search passages containing words that are in both, the question and the multiple-choice answer, instead of search passages containing words from the question and the answer, independently.

The second submodule (Information Retrieval- IR) was built using the Lucene[4] IR library. It is responsible for indexing the document collection, and for the further passage retrieval, given an hypothesis as a query.

The IR module returns a relevant passage for each hypothesis which is used as a support text to decide whether or not the hypothesis can be the right answer. For each hypothesis the first passage returned is taken (only one), which is considered the most important one. This process generates a pair "Hypothesis + Passage (*H-P*)", along with a lexical similarity score calculated by Lucene.

[2] http://wing.comp.nus.edu.sg/ qiu/NLPTools/JavaRAP.html
[3] http://nlp.stanford.edu/software/tagger.shtml
[4] http://lucene.apache.org/core/

3.3 Answer validation

Finally, the answer validation module aims to assign a score based on semantic similarity to the pair *H-P* generated in the *Information Retrieval* module. The reason for including this measure is that the lexical similarity score given by Lucene is not enough to capture the similarity between the hypothesis and the support text, when they do not share the same words. To overcome this problem, two things can be done: 1) To include a query expansion module trying to add synonyms, hyperonyms, etc, in order to obtain a higher lexical similarity, and 2) To add a semantic similarity algorithm which can discover the degree of similarity between two sentences, even though they do not share the same words exactly. For example in the hypothesis: "she esteems him is Annie Lennox's opinion about Nelson Mandela", the recovered passage is "Everyone one in the world respects Nelson Mandela, everyone reveres Nelson Mandela"; but the score assigned by Lucene is too low and it does not select that answer as the correct one. The addition of semantic similarity score will help to raise the score of these two phrases and select the correct answer because it will probably find the relation between the words "esteems", "revers" and "respect".

In order to determine whether or not the passage P is similar to an hypothesis H, we implemented an approach based in [8].

The similarity measure used in that paper [9] gives a weight to each word of the sentence in terms of the degree of specificity of the word. For example the words **catastrophe** and **disaster** gain more weight than words **could** and **should**. The similarity inter-words for both sentences is integrated into this measure. The two similarity measures proposed are: Corpus-based (PMI-IR) and Knowledge-based Measures (Wordnet[10]).

The similarity between two sentences *S1* y *S2* is given by the equation 1

$$sim(S_1, S_2) = \frac{1}{2}\left(\frac{\sum_{w \ \epsilon \ \{S_1\}} (maxSim(w, S_2) * idf(w))}{\sum_{w \ \epsilon \ \{S_1\}} idf(w)} + \frac{\sum_{w \ \epsilon \ \{S_2\}} (maxSim(w, S_1) * idf(w))}{\sum_{w \ \epsilon \ \{S_2\}} idf(w)}\right) \quad (1)$$

To find $maxSim$ we have used two semantic similarity measures between words, which are described as follows:

- **Mutual Information** *PMI-IR measure*. It comes from the pointwise mutual information formulae suggested by [11] as an unsupervised measure for the evaluation of semantic similarity of words. It is based on statistical data collected by an information retrieval engine over a very large corpus (i.e. the web). Given two words w_1 y w_2, its *PMI-IR* is measure by:

$$PMI - IR(w_1, w_2) = log_2 \frac{p(w_1 \& w_2)}{p(w_1) * p(w_2)} \quad (2)$$

- **WordNet measure**. It is based on the shortest path that connects two concepts in the taxonomy (hyperonyms, homonyms) extracted from Wordnet, a lexical database that groups the words in sets of synonyms called "synsets". The given score is in the interval 0 to 1, where the score 1 represents the equality of the concepts.

4 Experimental results

This section describes the data sets used for evaluating the methodology proposed in this paper. Additionally, the results obtained in the experiments carried out are reported and discussed.

In order to determine the performance of the system proposed in this paper we used the corpus provided in the QA4MRE task of the CLEF 2011. The features of the test data set is detailed in Table 1.

Table 1. Features of the test data set (QA4MRE 2011 task)

Features	**2011**
Topics	3
Topic details	Climate Change, Music & Society and AIDS
Reading tests (documents)	4
Questions per document	10
Multiple-choice answers per question	5
Total of questions	120
Total of answers	600

Table 2 presents the obtained results in terms of number of correct answered questions. It is shown that the semantic similarity measures are able to find some answers that otherwise with the lexical similarity measure are unable to find. The number of the different correct answers achieved by the PMI measure is 15 and the ones achieved by the Path measure is 8. The lexical similarity achieved 18 different correct answers, whereas the number of correct answers achieved by both, lexical and semantic similarity is 21. In total, the number of correct answers given by both similarity measures is 54 (45%). This precision overcomes the 32% achieved by the approach that uses only the lexical similarity measure.

Table 2. Comparison of the number of correct answers obtained with different similarity measures

Similarity Measures	**2011**
PMI	15
Path	8
Lexical	18
Both	21
Total (PMI + Lexical + Both)	54
Precision (Lexical + Both)	0.32%
Precision (PMI + Lexical + Both)	0.45%

5 Conclusion and Future Work

In this paper we have presented a methodology for tackling the problem of question answering for reading comprehension tests, making emphasis on the validation step. There were presented two semantic similarity measures, one based on PMI and the other one based on Wordnet, specifically the shortest path measure.

We have compared the performance of the system presented in this paper using the lexical and semantic similarity measures. We have observed that the semantic similarity measures are able to discover answers that with the lexical similarity measure could not be discovered.

As future work we would like to determine which question is more suitable to be validated by a semantic measure, and which one is better to be validated with a lexical measure. Making this process automatic will improve the overall precision of the methodology.

References

1. Hirschman, L., Gaizauskas, R.: Natural language question answering: the view from here. Nat. Lang. Eng. **7**(4) (December 2001) 275–300
2. Peñas, A., Hovy, E.H., Forner, P., Rodrigo, Á., Sutcliffe, R.F.E., Forascu, C., Sporleder, C.: Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation. In: CLEF. (2011)
3. Peñas, A., Hovy, E.H., Forner, P., Rodrigo, Á., Sutcliffe, R.F.E., Sporleder, C., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE at CLEF 2012: Question answering for machine reading evaluation. In: CLEF. (2012)
4. Hirschman, L., Light, M., Breck, E., Burger, J.D.: Deep Read: a reading comprehension system. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. ACL '99, Stroudsburg, PA, USA, Association for Computational Linguistics (1999) 325–332
5. Riloff, E., Thelen, M.: A rule-based question answering system for reading comprehension tests. In: Proceedings of the ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 13–19
6. Ng, H.T., Teo, L.H., Lai, J., Kwan, J.L.P.: A machine learning approach to answering questions for reading comprehension tests. In: In Proceedings of EMNLP/VLC-2000 at ACL-2000. (2000)
7. Vicedo, J.L., Ferrandez, A.: Importance of pronominal anaphora resolution in question answering systems. In: In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL. (2000) 555–562
8. Carrillo, M., Vilariño, D., Pinto, D., Tovar, M., León, S., Castillo, E.: FCC: Three approaches for semantic textual similarity. In: In proceedings of Semeval 2012, Montréal, Canada, Association for Computational Linguistics (2012) 631–634
9. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st national conference on Artificial intelligence - Volume 1. AAAI'06, AAAI Press (2006) 775–780
10. Miller, G.A.: Wordnet: a lexical database for the english language (1995)
11. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Comput. Linguist. **16**(1) (March 1990) 22–29